数据挖掘技术在教育中的应用研究

杨永斌
(重庆工商大学计算机科学与信息工程学院　重庆 400067)

摘　要　随着教育信息化进程的推进，产生并积累了大量的、复杂的数据，为了更充分、有效地利用这些数据，本文就数据挖掘技术在教育中的应用进行了一些探讨，并以教学评价作为简单的实例研究，目的在于发现大量教育数据中隐藏的、有用的知识，以指导教育、发展教育、为教育服务。
关键词　数据挖掘，建模，教育，教学，评价

The Application Research of Data Mining Technique in Education

YANG Yong Bin
(College of Computer Science and Information Engineer, Chongqing Technology and Business University, Chongqing 400067)

Abstract　With the developing of informationized education, a lot of comjplex data are produced. In order to more sufficient and effective use the data, this paper discusses the application of data mining in education and studies teaching examples in order to find useful and concealed education data to guide education, develop education and serve education.
Keywords　Data mining, Modeling, Education, Teaching estimation

1　前言

随着数据库技术的不断发展，数据库和数据仓库已经被广泛地应用于企业管理、产品销售、科学计算和信息服务等领域，数据量的不断增长对数据的存储、管理和分析提出了更高的要求，急需新一代的计算技术和工具，能够智能化地从大量的数据中提取出有用的信息和知识，于是数据挖掘技术应运而生。

数据挖掘技术从一开始就是面向应用的，诸如银行、电信、保险、交通、零售(如超级市场)等商业领域。数据挖掘所能解决的典型商业问题包括：数据库营销(Database Marketing)、客户群体划分(Customer Segmentation & Classification)、背景分析(Profile Analysis)、交叉销售(Cross selling)等市场分析行为，以及客户流失性分析(Churn Analysis)、客户信用记分(Credit Scoring)、欺诈发现(Fraud Detection)等等。随着教育信息化进程的推进，将数据挖掘技术应用于教育中，从大量的教育数据中发现隐藏的、有用的知识来指导教育、发展教育，成为当今势在必行的重要的研究课题。

2　数据挖掘的概念和技术

2.1　数据挖掘的概念

从技术角度看，数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。从商业应用角度看，数据挖掘是一种崭新的商业信息处理技术，也就是根据预定义的商业目标，对大量的企业数据进行探索和分析，揭示其中隐含的商业规律，并进一步将其模型化的先进有效技术过程。数据挖掘是一门交叉学科，它集成了许多学科中成熟的工具和技术，包括数据库技术、统计学、机器学习、模型识别、人工智能和神经网络等等。

2.2　数据挖掘中的常用建模技术

数据挖掘是如何工作的？数据挖掘工具是怎样准确地告诉你那些隐藏在数据库深处的重要信息呢？它们又是如何做出预测的？答案就是建模。建模实际上就是在你知道结果的情况下建立起一种模型，并且把这种模型应用到你所不知道的那种情况中。在数据挖掘中最常用的建模技术有：

2.2.1　统计　是涉及数据和描述的一个数学分支，它的主要任务就是了解已经收集到的有限数据，并从中作出关于潜在数据分布是什么的预测，其中回归技术是为创建预测模型而最广泛使用的一项统计技术。

2.2.2　近邻技术　是指为了预测在一个记录中的预测值是什么，在历史数据库中寻找有相似预测值的记录，并使用未分类记录中最接近的记录值作为预测值。对基础最近邻算法常常会作的改进是从 K 个最近的邻居中进行投票选择，而不是仅仅取决于距未知记录最近的邻居，这就是所谓的 K 近算法。

2.2.3　聚类　用于将记录聚集在一起，从而给出数据库的一个高层视图，有时也指数据分割，主要有两种类型的聚类方法：创建分层聚类和不分层聚类。

2.2.4　决策树　是指采取树的形式的预测模型，树的每个分支都是一个分类方法，树叶是带有分类的数据分割。最流行的生成树的算法有 CART(Classification and Regression Trees，分类回归树)和 CHAID(Chi Square Automatic Interaction Detector，X 检测法自动交感侦察器)。

2.2.5　人工神经网络　仿照生理神经网络结构的非线形预测模型，通过学习进行模式识别。

2.2.6　规则归纳　从统计意义上对数据中的“如果…那么…”规则进行寻找和推导。

3　数据挖掘环境及过程

3.1　数据挖掘环境

数据挖掘是指一个完整的过程，该过程从大型数据库中

杨永斌　副教授，主要研究方向：计算机算法、计算机网络、数据库技术、网络教育。

挖掘先前未知的，有效的，可实用的信息，并使用这些信息做出决策或丰富知识。数据挖掘环境可示意如图 1。

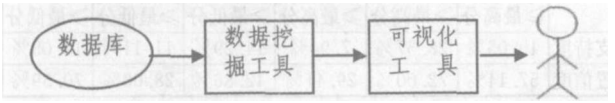

图 1　数据挖掘环境图

3.2　数据挖掘过程图

图 2 描述了数据挖掘的基本过程和主要步骤。

过程中各步骤的大体内容如下：

3.2.1　确定业务对象　清晰地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步，挖掘的最后结构是不可预测的，但要探索的问题应是有预见的，为了数据挖掘而数据挖掘则带有盲目性，是不会成功的。

3.2.2　数据准备

①数据的选择。搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据。

②数据的预处理。研究数据的质量，为进一步的分析作准备。并确定将要进行的挖掘操作的类型。

③数据的转换。将数据转换成一个分析模型。这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

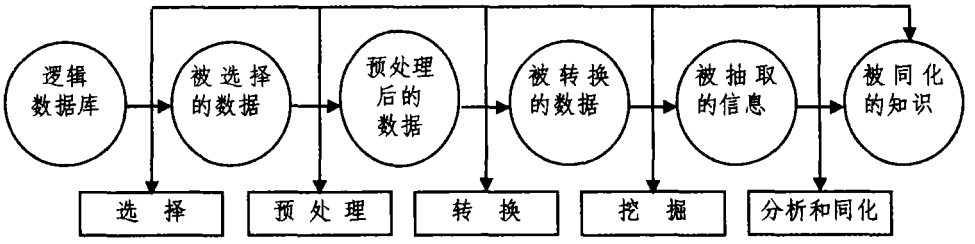

图 2　数据挖掘的基本过程和主要步骤

3.2.3　数据挖掘　对所得到的经过转换的数据进行挖掘。除了完善从选择合适的挖掘算法外，其余一切工作都能自动地完成。

3.2.4　结果分析　解释并评估结果，其使用的分析方法一般应作数据挖掘操作而定，通常会用到可视化技术。

3.2.5　知识的同化　将分析所得到的知识集成到业务信息系统的组织结构中去。

4　数据挖掘技术在教育中的应用

4.1　教师教学方法的选择

在教学过程中，教师可以采用多种教学方法来完成自己的教学任务：比如讲授法、讨论法、实验法、计算机辅助教学法、参观法、调查法、实习法等。在通常情况下，一般可以采取一种或几种方法进行。

据此可以用数据挖掘的方法来挖掘数据库中的数据，判定下一步我们应采取什么样的教学方法，以满足教学的需要，更利于学生对知识的吸收。如从每个学生对教学方法的评价以及不同的教学方法得出的教学成绩来进行分析，运用回归线性分析、关联规则的方法来判定此种教学方法适合哪一类学生或哪门课程，使得分层次教学能够得到更进一步的实施。

4.2　学习评价与学生特征挖掘

学习评价是教育工作者的重要职责之一。评定学生的学习行为既对学生起到信息反馈和激发学习动机的作用，又是检查课程计划、教学程序以至教学目的的手段，还是考查学生个别差异，便于因材施教的途径。评价要遵循“评价内容要全面、评价方式要多元化、评价次数要多次化、注重自评与互评的有机结合”的原则。利用数据挖掘工具，对学生的学习成绩、行为记录及奖励处罚数据库等进行分析处理，可以即时得到学生的评价结果，对学生出现的不良学习行为进行及时指正。同时可以减轻教师的工作量，还能够克服教师主观评价的不公正、不客观的弱点。

根据学生的基本信息、绩效信息、学习历史、学习偏好、知识结构等已有信息，挖掘学生特征，帮助学生修正自己的学习行为。通过对学生特征分析结果和事先制定的行为目标标准进行比较，教师能够帮助学生修正学习行为、提高学习能力、完善人格，有利于学生各方面素质的综合发展。

4.3　干预师生行为

学校教学管理数据库中记录着各届学生与教师的学习、工作、社会活动、奖励处罚等情况，利用数据挖掘的关联分析，寻找师生各种行为活动之间的内在联系。如“当存在 A，B 时可以推出 C”这样的规则，即当有 A 行为和 B 行为发生时，就会有 C 行为。在实际情境中，如果发现学生或教师已有 A，B 行为时，马上可以分析其产生 C 行为的可能性，并及时制定策略促进或制止 C 行为的发生。

也可以根据 A 学生和 B 学生在一起的行为，推导出 A 和 B 在一起是起着积极作用还是负面的影响，进而可以选定学生之间的关联，促进学生之间的互助行为。这样有助于培养整个班级乃至整个学校的精神风貌与学习氛围。

4.4　合理设置课程

在校学生的课程学习是循序渐进的，而且课程之间有一定的关联与先后顺序关系。在学习一门较高级课程之前必须先修一些先行课程，如果先行课程没有学好，势必会影响后继课程的学习。另外，同一年级学习同一课程的不同班级，由于授课教师、班级文化的不同，班内学生的总体成绩相差有时会很大。

利用学校教学数据库中存放的历届学生各门学科的考试成绩，结合数据挖掘的关联分析与时间序列分析等相关功能，就能从这些海量数据中挖掘出有用的信息，帮助分析这些数据之间的相关性、回归性等性质，得出一些具有价值的规则和信息，最终找到影响学生成绩的原因。在此基础上，对课程设置做出合理安排。

4.5　在考试方面

考试是对教和学效果的检验，是教学中必不可少的环节之一。虽然可以从分数角度总体评价在一定时期内取得的成绩，但并不能有效地说明分数的高低与哪些因素有关，无法知道教学中的成功与失败的关键因素，对教和学起不到促进作用。而且，学生分数的高低也与试题的质量有着密切的关系，因此，探索有效的方法来评价试题的质量（试题难易度、知

识点全面度等)在实际的教学过程中同样具有重要的意义。

如果将数据挖掘中的关联规则应用于试卷分析数据库中,然后根据学生得分情况分析出每道题的难易度、区分度、相关度等指标,教师就能够对试题的质量作出比较准确的评价,进而可以用来检查自己的教学情况及学生的掌握情况,并为今后的教学提供指导。

此外,数据挖掘技术还可用于网络教育资源建设、智能题库建设、招生就业管理、智能及个性化校园网建设等教育领域中。

5 实例研究

课堂教学评价不仅对教学起着调节、控制、指导和推动作用,而且有很强的导向性,是学校教学管理重要的组成部分,是评价教学工作成绩的主要手段。学校每学期都要搞课堂教学评价调查,积累了大量的数据。利用数据挖掘技术,将关联规则运用于教学评价数据中,就会挖掘出一些有用的数据,探讨教学效果的好坏与教师的年龄、职称之间有无必然的联系?课堂教学效果与教师整体素质关系如何?合理配置班级的上课教师,使学生能够较好的保持良好的学习态度,从而为教学部门提供了决策支持信息,促使更好地开展教学工作,提高教学质量。举例如下:针对不同教师所教学生的期末测试成绩统计如表 1 所示。

表 1 期末测试成绩统计表

	甲老师授课	乙老师授课	丙老师授课	合计
学生最高分	60	90	25	175
学生最低分	45	35	60	140
合计	105	125	85	315

不妨设"甲老师授课=> 学生得高分"为 X=> Y,则关联规则"甲老师授课=> 学生得高分"的支持度和置信度分别为:

Support(X=> Y)= 60/315= 19.05%

Confidence(X=> Y)= 60/105= 57.14%

依此类推,分别求出其他关联的支持度和置信度,如表 2

表 2 支持度与置信度表

	甲老师	乙老师	丙老师	甲老师	乙老师	丙老师
	≥最高分	≥最高分	≥最高分	≥最低分	≥最低分	≥最低分
支持度	19.05%	28.57%	7.94%	14.29%	11.11%	19.05%
置信度	57.14%	72.00%	29.41%	42.86%	28.00%	70.59%

研究其关联规则:"乙老师授课=> 学生得高分"的支持度和置信度分别为 28.57%和 72%,"丙老师授课=> 学生得低分"的支持度和置信度分别为 19.05%和 70.59%,表明这两条规则在很大程度上是成立。

当然,在进行教学评价时,仅根据学生成绩来反映教学效果的好坏是比较片面的,还应综合考虑其他因素,如教师的年龄、职称及学生各项素质指标等,及时地对教师的教学、学生的学习提供指导。

结束语 学校多年的教学和管理工作积累了大量的数据,并且随着教育信息化的逐步实施、完善和发展,数据信息不断地增长,但这些数据还未得到有效利用,是一个待开发的宝藏。将数据挖掘技术应用到教育中,从中发现有用知识,不仅可以促进教育体制上进一步改革、发展和完善,而且能够比较客观地反映教育中存在的问题,对学校教学管理的决策支持是十分有意义的,但此类研究目前国内尚不多见,需要更多的研究者投入到此研究领域,以取得技术和应用上的突破。

参 考 文 献

1 杨永斌. 数据挖掘技术在证券业中的应用研究[J]. 重庆工商大学学报(自然科学版), 2005(5): 461~463
2 (美) Kantardzic M 著. 数据挖掘——概念、模型、方法和算法[M]. 闪四清, 等译. 北京: 清华大学出版社, 2003
3 陈京民, 著. 数据仓库与数据挖掘技术[M]. 北京: 电子工业出版社, 2002
4 (美) Berson A, Smith S, Thearling K 著. 构建面向 CRM 的数据挖掘应用[M]. 贺奇, 等译. 北京: 人民邮电出版社, 2001
5 (美) Berry M J A 著. 高管商学院. 数据挖掘(全球顶尖商学院 MBA 课程精华)[M]. 袁卫, 等译. 北京: 中国劳动社会保障出版社, 2004
6 (美) Linoff G S, Berry M J A 著. Web 数据挖掘: 将客户数据转化为客户价值[M]. 沈钧毅, 等译. 北京: 电子工业出版社, 2004

(上接第 127 页)

5.5 特征添加和删除

由于空间数据具有时间动态性,空间数据会随着时间的变化而改变。如在地图上添加或删除点状、面状或线状特征。以增加和删除点状房屋特征为例,关键代码如下:

1) 特征增加

```
If Recs. Updatable Then' 添加特征
With layer. Records
   . Edit
   . Fields("Shape"). Value = P1
   . Fields("Layer"). Value = "房屋"
   ……
     . Update
     . StopEditing
   End With
  Map1. Refresh
End If
```

2) 特征删除

```
With layer. Records
       . Edit
       . Delete
         . MoveNext
         . StopEditing
End With
Map1. Refresh
```

此外利用 MO 可实现空间分析、图层渲染、图层的修改、地图投影变换、制图输出等功能,在此不再叙述。

结束语 本文以 ArcGIS 为平台对已有的资源 AutoCAD(.dwg)格式的数据,利用 Geodatabase 数据模型通过 ArcSDE 导入商业数据库,实现了图文一体化转换和存储,此外,ArcInfo 平台也可将 Shapefile 格式和 TAB 格式的数据通过 ArcSDE 导入数据库中。通过 Geodatabase+ ArcSDE 存取 GIS 空间数据的方法将 GIS 中存在的数据内容、数据模型和数据格式多样性等问题转移到数据库领域中,为最终实现空间数据共享提供了一种解决思路。

参 考 文 献

1 隋铭明, 李宗华, 等. 空间数据共享与互操作在规划国土管理部门的实现初探[J]. 地理空间信息, 2005, 3: 16~18
2 孟华, 李晓东, 韩敏, 等. 基于 Geodatabase 和 ArcSDE 的湿地 GIS 数据库技术研究与应用实例[J]. 计算机应用研究, 2005(10): 184~187
3 ArcSDE SQL Server Administrator Lecture Book. 2001. 96~98
4 黄林进, 甘雪梅. 城市地理信息系统空间基础数据建设探讨[J]. 现代测绘, 2003(4): 38~39
5 周小成, 焦道振. 基于 Geodatabase 的 CAD 数据到 GIS 的解决方法[J]. 现代测绘. 2004, 27: 15~17